sustainability

MDPI

Article

Construction of Meteorological Simulation Knowledge Graph Based on Deep Learning Method

Ziwei Xiao and Chunxiao Zhang *

School of Information Engineering, China University of Geosciences, Beijing 100083, China; cugbxzw@163.com
* Correspondence: zcx@cugb.edu.cn; Tel.: +86-010-82321744

Abstract: With the maturity of meteorological simulation technology, the research literature in this field is undergoing a rapid increase. The published literature can provide useful guidance for current research to get scientific results; however, it tends to be rather time consuming to obtain exact knowledge from massive literature, and it is necessary to transform the literature into structured knowledge to meet the efficient management, sharing, and reuse of meteorological simulation knowledge. In this paper, methods of meteorological simulation knowledge extraction and knowledge graph construction are proposed. A deep learning model based on bilateral long short-term memory-conditional random field (BiLSTM-CRF) is used to extract the meteorological simulation knowledge from the massive literature. Then, the Neo4j graph database is used to construct the meteorological simulation knowledge graph. Based on the meteorological simulation knowledge graph, it can realize the structured storage and integration of meteorological simulation knowledge, which can bridge the gap in the transformation of massive literature to sharable and reusable knowledge. Furthermore, the meteorological simulation knowledge graph can be used as an expert resource and contribute to sustainable guidance and optimization for meteorological simulation research.

Keywords: meteorological simulation; knowledge graph; knowledge extraction; deep learning; bilateral long short-term memory (BiLSTM); conditional random field (CRF)

Citation: Xiao, Z.; Zhang, C. Construction of Meteorological Simulation Knowledge Graph Based on Deep Learning Method. *Sustainability* **2021**, *13*, 1311. https://doi.org/10.3390/su13031311

Received: 30 December 2020
Accepted: 23 January 2021
Published: 27 January 2021

1. Introduction

Meteorological processes have an extremely close relation with social development and economic growth, and they play a significant role in natural disaster prevention, air pollution control, ecological environment protection, energy structure optimization, and green gas issues [1–5]. With the transformation of geographical research from static surface environment to dynamic geographical processes, as an essential part of geographical processes, many related fields have carried out research on meteorological simulations. Due to the complexity of meteorological processes, the research involves comprehensive multidisciplinary knowledge, including geography, meteorological dynamics and environmental science, which requires multifield expertise to obtain an accurate and scientific simulation result [6,7]. However, in current meteorological simulation research, accumulated research papers and other related literature have not been sufficiently converted into structured and reusable knowledge, and the expertise is inconvenient to obtain. It is hard to efficiently share and reuse the meteorological simulation knowledge distributed in the literature content, and the guidance of conducted research results for current research cannot be fully utilized. Therefore, the extraction of meteorological simulation knowledge and construction of knowledge graph are recommended to transform the unstructured literature context into structured knowledge; it can enhance the ability of knowledge management, sharing, and reuse; furthermore, it can provide scientific knowledge support for meteorological simulation research.

In the era of big data, information and data have increased dramatically, and people have found that it is rather inefficient to obtain useful information and knowledge from the massive and noisy plaintext. Therefore, it is vital to improve the transformation ability of "data-to-knowledge" [8,9], which prompts the development of a knowledge graph. Knowledge graph technology based on natural language processing provides a method for extracting structured knowledge from massive texts and has attracted considerable attention in knowledge engineering research. A knowledge graph is a large-scale semantic network that can realize the structured storage of complex interconnected data. It is composed of nodes and directed edges, which can express the knowledge (including concepts, entities, and attributes) and their semantic relationships explicitly. Through graph structure, the management and association analysis of knowledge can be quite efficient. Knowledge graph technology has promoted knowledge interconnection, and research in knowledge graph construction, knowledge reasoning, and knowledge applications has gradually formed a knowledge service in the era of big data [10–13]. Due to the excellent capability of knowledge graph technology to manage complex and interconnected knowledge, research on knowledge graph technology has been carried out in many professional fields, and it has dramatically improved the management and reuse ability of knowledge in these fields. However, because of the scarcity of domain corpora and unavailable knowledge extraction methods oriented to meteorological simulation literature, the general methods of knowledge graph construction are not adaptive in the meteorological simulation field.

In this paper, meteorological simulation knowledge extraction and knowledge graph construction methods are proposed to enhance the knowledge management, sharing, and reuse levels in the meteorological simulation field. The main contributions of this study can be summarized as follows:

(1). A knowledge extraction model based on BiLSTM-CRF is proposed, which can recognize and extract structured knowledge from the unstructured literature content in the meteorological simulation field.
(2). Through the construction of a meteorological simulation knowledge graph, the structured storage of meteorological simulation knowledge is realized.
(3). Efficient knowledge management, knowledge retrieval, and association analysis methods based on the meteorological simulation knowledge graph are demonstrated.

The following content of this paper consists of five sections. The related work of meteorological simulation and knowledge graph technology is introduced in Section 2. Then, Section 3 introduces the methodology of knowledge graph construction. Subsequently, Section 4 is a case study of meteorological simulation knowledge graph construction. In Sections 5 and 6, the research is discussed and summarized, and future work directions are proposed.

2. Related Work

2.1. Meteorological Simulation

With the development of computer science, geographic information system, and remote sensing technology, model-based meteorological numerical simulation experiments are gradually emerging. Overall, the research can be divided into two main aspects. First, the research about interactions between meteorological and environmental factors, such as underlying surface properties, terrain, urban landscape pattern, and energy structure [14–18], has been widely conducted. The other aspect is the impact of the meteorological environment on air pollution, extreme weather and catastrophe, and vegetation growth [19–23]. As in the related research listed above, typically, the main method of current research on meteorological simulation is to use models to carry out numerical simulation research. Scholars have conducted meteorological simulation analysis from a multiscale perspective, and the research achievements are fruitful.

Due to the complexity of the meteorological process, the simulation model design and parameter selection have a significant impact on the accuracy of the numerical simulation results. To avoid scientific losses caused by inappropriate model and parameters

selection, the knowledge in the published literature can provide significant guidance for current research; however, the literature is undergoing an explosive increase, and it is time consuming to find proper literature and locate the needed knowledge. Thus, it is necessary to organize and manage the knowledge of models, parameters, data, and result evaluations in the large amount of published research literature as well as improve the sharing and reuse of meteorological simulation knowledge.

2.2. Knowledge Graph

Knowledge graph technology originated from semantic networks in the 1960s [24]. In the 1990s, the idea of "ontology" was introduced into the knowledge representation methods [25]. With the development of the Worldwide Web and open link data, Google officially proposed the concept of the knowledge graph in May 2012. With the rapid development of big data, cloud computing, and artificial intelligence, knowledge service research represented by the knowledge graph is in the ascendant [26–28]. As a result of the excellent knowledge storage and management capability, the knowledge graph has become a key part of big data analysis.

Generally, knowledge graphs can be divided into two types: open filed knowledge graphs and domain knowledge graphs. The current well-known knowledge graphs in open fields include Freebase [29], DBpedia [30], Wikidata [31], YAGO2 [32], Zhishi.me, and OpenKG.CN. In addition, in the specific domain, research on knowledge graph construction and applications has also been conducted. Yan et al. [33] used the edit distance algorithm and latent Dirichlet allocation (LDA) algorithm to construct a water-affair knowledge graph and an information recommendation system to enhance water-affair data integration and application. Rotmensch et al. [34] explored an automatic process to learn and extract knowledge from massive electronic medical records and construct a health knowledge graph, which can provide scientific support for medical diagnoses and clinical decisions. Ho et al. [35] proposed a novel method for sharable and structured design rule construction in additive manufacturing based on machine learning and knowledge graph construction, which accelerated innovations in decision making in the additive manufacturing field. Heck et al. [36] built a deep structured semantic model of the knowledge graph embedding method to learn all the concepts of Wikipedia, which can improve the semantic relatedness computing ability of parsers and realize an effective error reduction in the semantic parsing of Twitter dialogues. These studies have realized an efficient acquisition and management of knowledge as well as prompt intellectualization in the corresponding field.

With the introduction of high-performance computing and artificial intelligence into the geography field, knowledge services have become the goal of geographic information science research [37]. The geographical knowledge graph is considered to be the key to extending traditional geographic information services to geographic knowledge services [38], and it has become a popular issue in geographic information science research. Scholars have begun to research the geographic knowledge graph, geographic knowledge engineering, and geographic knowledge services. Xu et al. [39] employed the conditional random field and multichannel convolutional neural network and achieved excellent results in the named entity recognition of geographic subjects. Guo et al. [40] proposed a region geographical knowledge graph construction method based on geographical ontology and developed a search system to meet the knowledge retrieval requirements. Wang et al. [41] realized an automatic annotation technology based on encyclopedia knowledge, which can generate high-quality corpus for machine learning model training in the task of geographical relation extraction. Zhang et al. [42] constructed a personalized virtual landslide disaster environment based on the deep learning and knowledge graph method, and they developed a landslide disaster scene data recommendation mechanism for multilevel users. Shi et al. [43] detected the meteorological events in the social network information by a wide-grained capsule network, which is an attempt to extract the meteorological knowledge entity in network text. Overall, in the field of meteorological simulation, the maturity of simulation technology has brought a large number of meteorological simulation research

results. However, due to the limitation of the ability to convert "data" to "knowledge", the accumulated research literature has not been fully transformed into structured knowledge that can be shared and reused. It is necessary to extract the knowledge from the massive literature and realize the structured storage in the knowledge graph to enhance knowledge management and accessibility for researchers in the meteorological simulation field. In the field of knowledge graph research, knowledge extraction from plaintext, such as literature content, usually requires natural language processing and deep learning methods. Additionally, because of the complexity of meteorological processes and strong professional requirements in simulation research, there is no sufficient corpus and feasible knowledge extraction method. The current method is not fully applicable in the meteorological simulation field, which is an obstacle to meteorological simulation knowledge graph construction. Research on a meteorological simulation knowledge graph is still rare, and it is difficult to conduct knowledge services.

Focusing on the above-mentioned problems and deficiencies in the current related research, the methods of meteorological simulation knowledge graph construction based on deep learning are proposed; the proposed methods concentrate on the knowledge extraction from massive unstructured meteorological simulation literature and the structured knowledge storage in knowledge graph. Based on the knowledge graph, the management, sharing, and reuse ability of meteorological simulation knowledge can be enhanced dramatically. Users can obtain structured knowledge related to their research efficiently from the knowledge graph, rather than looking up the massive and noisy literature to find the knowledge they need; knowledge guidance in current research can be sufficiently exerted, and consequently, the accuracy and scientificity of meteorological simulation can be improved. Ultimately, it can enhance the support abilities of meteorological simulation research in policies establishment of regional meteorological issues [44–46].

3. Methodology

As shown in Figure 1, the research framework of meteorological simulation knowledge extraction and knowledge graph construction consists of three main components: preparation of the knowledge source, acquisition of meteorological simulation knowledge, and construction of meteorological knowledge graph. In the first component, the meteorological simulation literature published online was collected, the semi-structured information on the web pages and the literature content was selected as the knowledge source, and a meteorological simulation ontology library was constructed. The second component is the key part of this research, which includes the literature information acquisition from semi-structured web pages by web crawler and knowledge extraction from unstructured literature content by a bilateral long short-term memory-conditional random field (BiLSTM-CRF) model. Subsequently, in component 3, the fusion of knowledge from diverse sources was conducted, and the meteorological simulation knowledge was ultimately constructed.

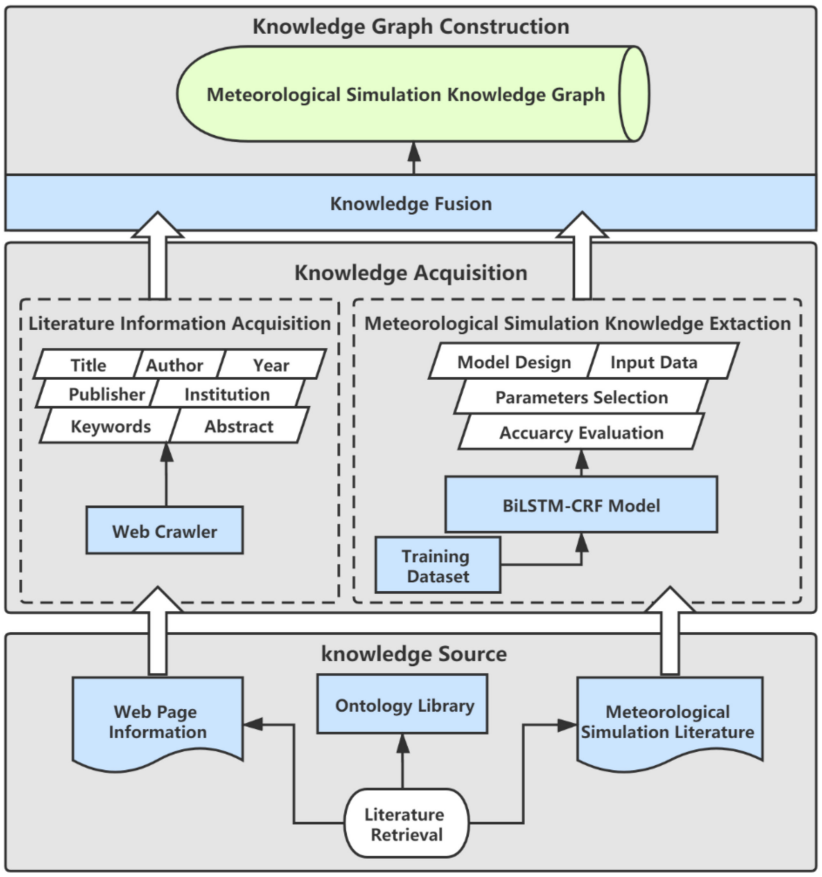

Figure 1. Framework of meteorological simulation knowledge extraction and knowledge graph construction.

3.1. Preparation of Knowledge Source and Construction of Ontology Library

China national knowledge infrastructure (CNKI) is one of the most famous and authoritative knowledge publishing platforms in China, and extensive advanced research results have been published on it. Thus, in this research, the meteorological simulation literature published on CNKI was selected as the knowledge source. First, the literature related to meteorological simulation was searched and collected through preprocessing such as text format conversion and paragraph screening, the part describes the simulation model design, and result accuracy evaluation was selected and stored to construct a corpus database, which can be used as a knowledge source for knowledge graph construction. In addition to the knowledge in the literature content, the title, authors, institutions, publisher, and keywords are also important knowledge of meteorological simulation research. Therefore, this information was collected and added into the knowledge graph in the subsequent processes.

Ontology refers to a clear, formal, and standardized description of concepts and their relationships in specific field [25,47]. It is substantially a top level of the knowledge graph, and it describes the data pattern in the graph. The pattern layer of the knowledge graph is constructed through the design of the ontology library, which defines a conceptual hierarchy with a clear structure of the knowledge and semantic relations in the graph. The pattern layer of the meteorological simulation knowledge graph is mainly composed of 6 core elements, and the ontology is expressed as:

Ontology = {LiteratureInformation, InputData, SimulationScope, ParameterScheme, SimulationTime, ResultVilidation, Relation}.

LiteratureInformation refers to the title, authors, institutions, publisher, and keywords of the literature. InputData refers to the data used in the simulation experiment, such as field data and underlying surface data. SimulationScope refers to the study area location,

simulation scales, resolution setting, simulation area division, and nested grid. ParameterScheme refers to the parametric scheme of the simulation model. SimulationTime refers to the time settings of the simulation process, and ResultVilidation refers to the evaluations of the simulation result, such as error assessment and correlation coefficient. Relation refers to the semantic relationship of these ontologies. The ontology library defines the data pattern in the process of knowledge acquisition and graph construction.

3.2. Meteorological Simulation Knowledge Acquisition

Under the guidance of domain ontology, the data layer of the knowledge graph is constructed. The literature information, such as the author, institutions, and key words are a kind of important knowledge of the meteorological simulation research; they are always displayed in the info box of the literature web pages in a semi-structured form. Hence, in this research, the web crawler was employed to parse the HTML codes of the web pages to obtain the literature information and store them in the structured data sheet for knowledge graph construction.

Regarding the simulation knowledge in unstructured literature content, a knowledge extraction model based on BiLSTM-CRF was used to recognize and extract the knowledge. The structure of the model is shown in Figure 2. It contains 5 layers, including an input layer, embedding layer, BiLSTM layer, CRF layer, and output layer.

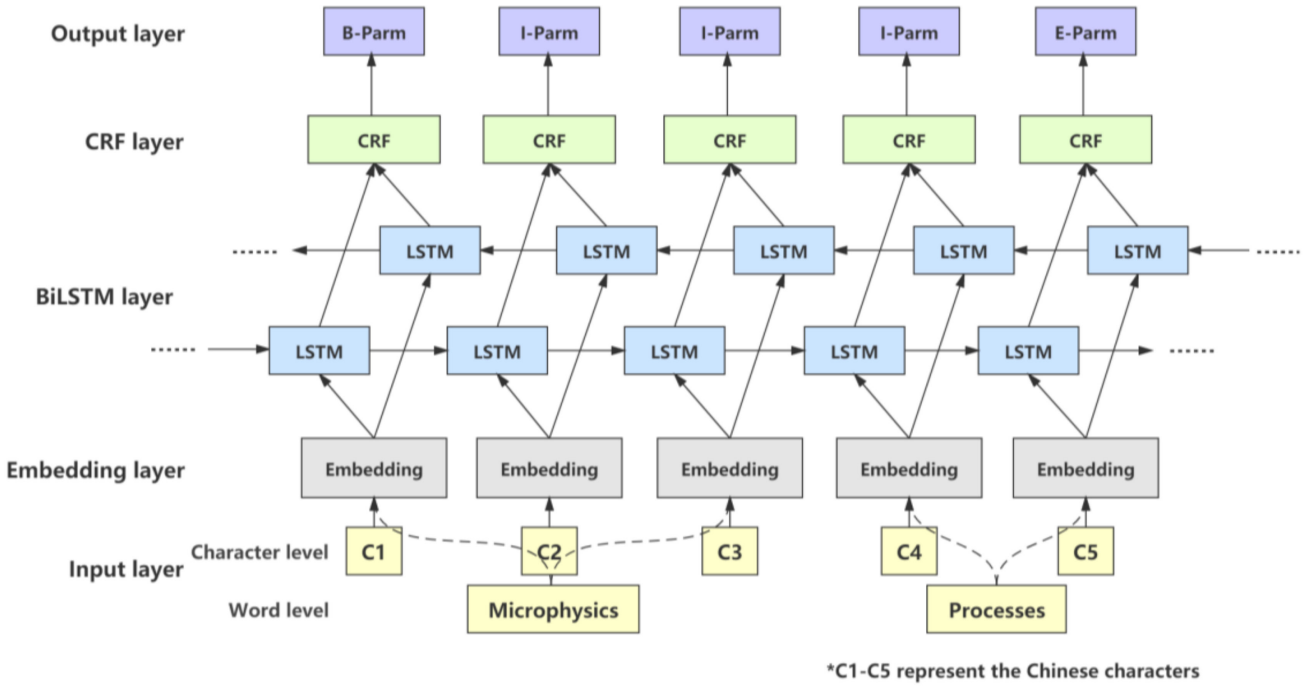

Figure 2. Meteorological simulation knowledge extraction processes.

3.2.1. Input Layer and Embedding Layer

The corpus was input into the model sentence by sentence. The pure character-level input lacks word-level information; likewise, the pure word-level input is very dependent on the accuracy of text segmentation, and word segmentation errors will cause entity boundary recognition errors. Therefore, in the input layer, the character-level and word-level information was integrated and input into the model, which can make full use of the context information.

The natural language cannot be processed by the BiLSTM directly; the semantic information in the text needs to be digitized for further processing. Hence, in the embedding layer, the input words and characters are transformed into eigenvectors by word embedding. Word embedding is a language processing model trained by corpora, such as the

continuous bag-of-words (CBOW) and Skip-Gram (SG) model. Based on the token in the corpus and its context, the word embedding model can map each token into the numeric vector space and generate an eigenvector with semantic information [48]. Word2vec, a word embedding tool provided by Google, was employed in this study to convert the word and character inputs into eigenvectors, which can be processed by the subsequent BiLSTM-CRF model.

3.2.2. BiLSTM-CRF Model

Since the knowledge in meteorological simulation literature is distributed in plaintext, natural language processing technology is required to process the semantic information and recognize the knowledge entity. The BiLSTM-CRF model is a natural language processing model based on deep learning, which is widely used in knowledge recognition and extraction; it has an excellent performance in unstructured text processing [49]. Knowledge recognition and extraction is essentially a sequence labeling task, an input sequence of sentence is processed by the BiLSTM-CRF model and it outputs a label sequence, which labels the knowledge entities in the sentences. The BiLSTM-CRF model is composed of two layers of long short-term memory (LSTM) with forward and backward directions, and a layer of conditional random field (CRF).

LSTM is a modification of the recurrent neural network (RNN); it expands the computing unit of the traditional RNN and uses a gating mechanism, which can solve the problem of gradient disappearance or explosion in the processing of long sequence data. The structure of the LSTM unit is shown in Figure 3.

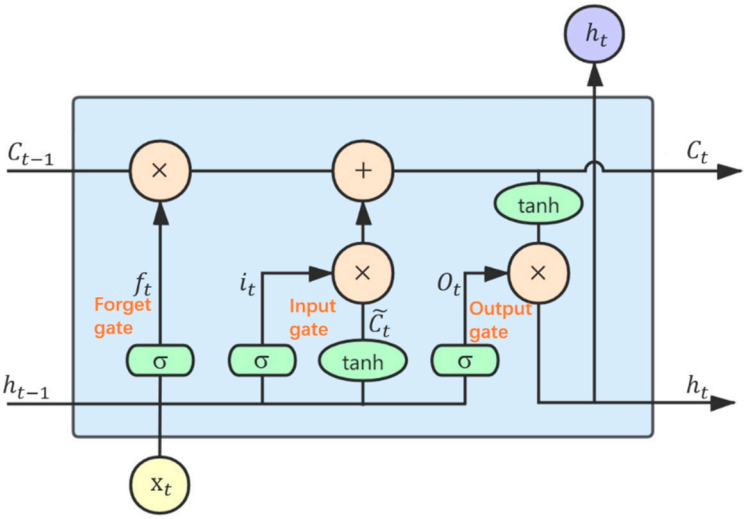

Figure 3. Long short-term memory (LSTM) internal structure.

In sentences processes of LSTM, through the gating mechanism, it is can selectively change the update and retention in the processing of the information data stream [50]. The formulas of the gating mechanism are expressed as follows:

$$\begin{aligned} f_t &= \sigma\left(w_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \\ i_t &= \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \\ \widetilde{C}_t &= \tanh(w_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \\ O_t &= \sigma(w_O \cdot [h_{t-1}, x_t] + b_O) \\ h_t &= O_t \cdot \tanh(C_t). \end{aligned} \tag{1}$$

In the formula, h_{t-1} is the output of the hidden layer at the previous moment, X_t is the input at the current moment, σ and tan h are the activation functions, w is the weight matrix, and b is the bias vector. f_t, i_t, and O_t are the outputs of the forget gate, input gate, and output gate, respectively. $\widetilde{C}_t$ is the information to be added to the cell state, C_t is the updated cell state at the current moment, and h_t is the final output result of the current LSTM unit [51].

In the text processes, the forget layer can analyze the current input and delete the useless information of the former text. The input layer controls the useful information that can be added into the information stream. Then, the results of the forget layer and input layer are processed and integrated in the output layer, which is the final output of the current token. LSTM can control the long-distance retention of effective information and the rapid forgetting of useless information, which is very suitable for processing long text sequences. BiLSTM is composed of a forward LSTM and a backward LSTM; in text processing, the text can be processed from both front and back directions, and the semantic information of the context can be fully considered.

However, in natural language, there are some syntactic constrains in the sentences, which cannot be captured by BiLSTM precisely; only use of the softmax function in its output layer may lead to prediction errors. Therefore, CRF is used to compensate for the shortcomings of BiLSTM in syntactic feature processing [52]. CRF is a discriminative probability model, which is a sequence processing algorithm based on hidden Markov model (HMM) promotion. It can receive an input sequence such as $X = \{x_1, x_2, x_3, x_4, \ldots, x_n\}$ and output the target sequence $Y = \{y_1, y_2, y_3, y_4, \ldots, y_n\}$, which can predict the conditional probability of output variables by input variables while considering the relationship between the front and back position features of each output [53]. Hence, in this study, the outputs of the BiLSTM layer are fed into a CRF layer, which is based on the BiLSTM-CRF model, while the semantic information and syntactical constrains in meteorological simulation literature can be fully processed to realize more accurate extraction of meteorological simulation knowledge.

The structure of the BiLSTM-CRF model is shown in Figure 4. The meteorological simulation literature text is segmented into token sequences and fed into the model; when the BiLSTM-CRF model processes the input sequence $X = \{x_1, x_2, x_3, x_4, \ldots, x_n\}$, for the input x_t, the forward LSTM calculates the context eigenvectors $\overrightarrow{h_t}$ before x_t, and the backward LSTM calculates the context eigenvectors $\overleftarrow{h_t}$ after x_t. Splicing the calculation results in the forward and backward LSTMs as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$, which is the complete semantic eigenvectors representation of the input token x_t. The input eigenvector h_t is trained by the BiLSTM network to obtain the label prediction probability matrix p_t in the sequence. Then, p_t is input into the CRF layer, and the state transition matrix A_t between the front and back labels is calculated in the CRF layer. After processing by two layers, the final label prediction result is $Y = \{y_1, y_2, y_3, y_4, \ldots, y_n\}$, and the probability of prediction result is:

$$p(X|Y) = \frac{e^{s(X,Y)}}{\sum_{\overline{y} \in Y_x} e^{s(X,\overline{y})}}. \quad (2)$$

Y_x is all the possible labels of the input X, $\overline{y}$ refers to the right label of X, $s(X,Y)$ represents the probability of the prediction, and $s(X,Y)$ is defined as:

$$s(X,Y) = \sum_{i=0}^{n} A_{y_i,y_j} + \sum_{i=0}^{n} P_{i,y_j}. \quad (3)$$

P_{i,y_j} represents the probability that the X in position i is predicted as label y_j. A_{y_i,y_j} represents the probability of transition from y_i to y_j in the state transition matrix calculated by the CRF layer. Take the logarithm of both sides of $p(X|Y)$ to obtain the maximum likelihood function of sequence prediction as:

$$\ln(p(X|Y)) = s(X,Y) - \ln(\sum_{\overline{y} \in Y_x} s(X,\overline{y}))$$

$$Y^* = \underset{\overline{y} \in Y_x}{\arg max}\, s(X, \overline{y}). \quad (4)$$

Select the label with the largest prediction score Y^* as the final label sequence prediction result.

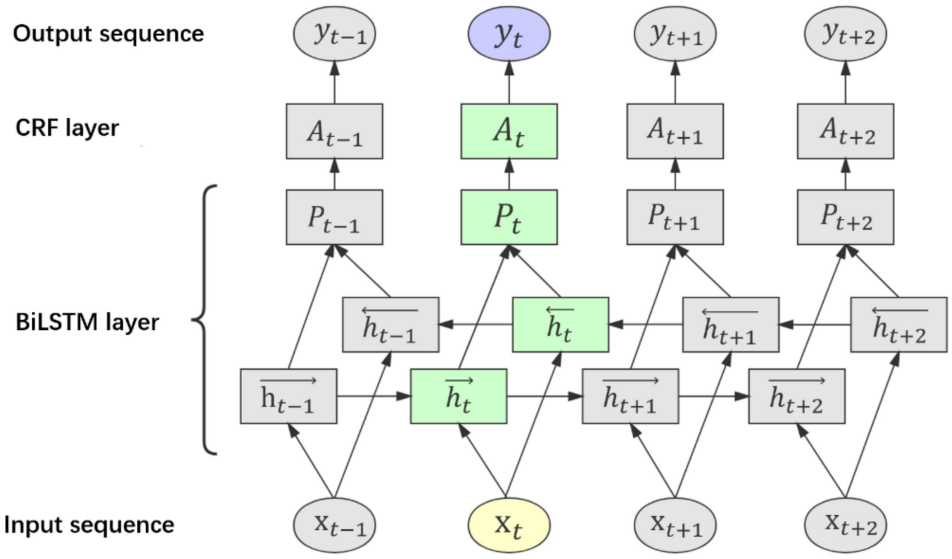

Figure 4. Bilateral long short-term memory-conditional random field (BiLSTM-CRF) model structure.

3.2.3. Output Layer

The input eigenvector of each token was processed by the BiLSTM-CRF model. Finally, the prediction label of each token in the sentences can be obtained in the output layer. The labels of each token indicate the type of knowledge entities; according to them, the specific entities, attributes, and concepts of meteorological simulation knowledge can be recognized and extracted. For the relationships between knowledge, the relationship between entities can be associated by domain prior knowledge, and the relationship of attributes and concepts can be associated by attribute names and values or locations in the sentences. For example, if two entities have labels of attribute name and attribute value, respectively, and they are in one sentence, they are considered to have a semantic relationship and will be associated in the knowledge graph.

3.2.4. Construction of Training Dataset

As the BiLSTM-CRF model is a supervised learning method, a large-scale training corpus dataset is required to train the model; however, the meteorological simulation is a professional field, the corpus in open fields is not applicable for meteorological simulation knowledge extraction, and there is no ready training dataset that can be used to train this model. Hence, a training dataset generation method by manual annotation and data augmentation is proposed. First, the corpus annotation scheme was defined based on the study of meteorological simulation theories, ontology, and expert guidance. The entities and attributes defined in the ontology library, including SimulationScope, InputData, ParameterScheme, SimulationTime, and ResultValidation, were annotated by the BIOES annotation scheme ('B-entity' refers to the beginning word of an entity, 'I-entity' refers to the intermediate word, 'E-entity' refers to the end word, 'S-entity' refers to a single word as an entity, and 'O' refers to other nonentity words). The annotation scheme is shown in Table 1.

Based on the annotation scheme, the corpus was manually annotated, and each character in the corpus was regarded as a token and matched to a corresponding label. Figure 5 shows the annotation for a sample sentence: "The center of simulation area is 30° N,100° E. The WSM5 scheme was selected as the microphysical parameterization scheme (in Chinese)." As the exported annotation shows, a complete knowledge entity is composed

of a B-entity label in the beginning, several I-entity labels in the middle, and a E-entity label in the end sequentially.

Table 1. Corpus annotation scheme.

Type	Instances	Label
SimulationScope	Resolution, nested grids, latitude, longitude, grids number, etc.	Area/AreaV [1]
InputData	Initial field, lateral boundary condition, terrain data, land use data, etc.	Data/DataV
ParameterScheme	Microphysical parameterization scheme, cumulus convective scheme, land surface scheme, radiative scheme, etc.	Parm/ParmV
SimulationTime	Integration time, integration step, start/end time, output interval, etc.	Time/TimeV
ResultValidation	Average error, absolute average error, root mean squared error, correlation coefficient, etc.	Val/ValV

[1] Area/AreaV refers to the attribute name and value of the instance, respectively, and the other labels are the same as above.

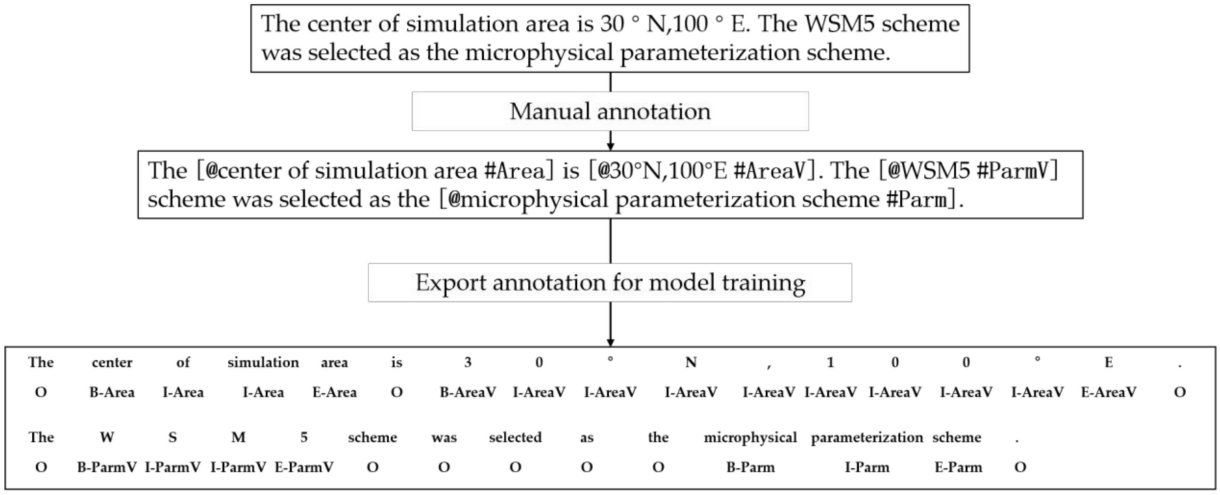

Figure 5. Sample of corpus annotation.

However, the construction of datasets by manual annotation is so dependent on the professional knowledge of annotators that it is time consuming and labor intensive. In this study, a small-scale training dataset was constructed by manual labeling under the guidance of experts, and the dataset was expanded by the data augmentation method to obtain a large-scale dataset that complied with the requirement of model training [54]. The data augmentation methods proposed in the article include the following:

- Choose n words randomly and replace them with their synonyms
- Choose n words randomly and insert their synonyms behind them
- Choose n words randomly and change their positions
- Choose n words randomly and delete them.

According to the recommended parameters in this article, the hand-crafted dataset was augmented and expanded to 10 times the size, which can be used for BiLSTM-CRF model training. The article claimed that the augmentation method above will not cause obvious semantic changes in the corpus. Additionally, some man-made noise was introduced into the dataset, which can improve the generalization ability of the model. In the case of an insufficient corpus, using data augmentation can significantly improve the model training result.

3.3. Construction of Meteorological Simulation Knowledge Graph

3.3.1. Knowledge Fusion

The corpus of this study was mostly Chinese literature. Due to the diversity of Chinese descriptions, there may be multiple expressions of the same knowledge entities. For the extracted knowledge, the different expressions of the same knowledge are supposed to be fused and unified; otherwise, it may cause data redundancy in knowledge storage. Therefore, it is necessary to define a suitable similarity measurement and use clustering or threshold setting methods for knowledge fusion.

In the process of knowledge extraction, each token of sentences is transformed into an eigenvector by word embedding, which can represent the semantic information of every token. Hence, the same method is used in the knowledge fusion process to measure the similarity among knowledge. Word segmentation is performed on the identified entity, attributes, and concepts, the word frequency of the word segmentation is calculated, and the knowledge entities are transformed into semantic eigenvectors, the cosine values of the angle between the vectors are calculated, and a greater cosine value of the angle means higher semantic similarity. By setting the semantic similarity threshold, the entities whose semantic similarity is greater than the threshold are considered to be the same knowledge entity, and they are fused.

3.3.2. Knowledge Storage and Knowledge Graph Construction

Through the above processes, the information and data of different forms are transformed into structured knowledge entities and relations. For meteorological simulation knowledge with clear structures and abundant attributes and relationships, the storage method of graph databases has obvious advantages [55–57], which can display the knowledge of meteorological simulation from multiple dimensions such as entities, attributes, and concepts. In addition, it is convenient to use graph query language and graph mining algorithms to carry out relationship extension calculations and specific applications of knowledge graph. Neo4j is a high-performance graph database that has superior capability in processing billions of entities nodes and relation edges; consequently, it is widely used in knowledge graph construction [58]. In this study, Neo4j was chosen as the storage of the knowledge graph, and the knowledge was stored in the following format, as shown in Table 2.

Table 2. Storage format of the meteorological simulation knowledge graph.

Knowledge Format	Neo4j Storage Format
[entity, relation, entity]	entity node—relation—›entity node
[entity, attribute]	entity node—attribute relation—›attribute node
[entity, attribute, attribute value]	entity node—attribute relation—›attribute node: value

The entities and attributes were stored as nodes, and the semantic relations were stored as edges to realize the mapping of structured knowledge to the knowledge graph.

4. Case Study

4.1. Meterial Preparation and Knowledge Extraction

The weather research forecast (WRF) model is a new generation of mesoscale forecasting model and assimilation system jointly developed by the American meteorological community. It has great performance and is widely used in various meteorological process studies. This study uses the meteorological simulation by the WRF model as an example to construct a meteorological simulation knowledge graph.

According to the knowledge graph construction method proposed in this paper, first, we constructed the pattern layer of the meteorological simulation knowledge graph by establishing a comprehensive and clear WRF model ontology library. Then, to build the data layer under the guidance of ontology, we searched the meteorological simulation

literature on CNKI with "WRF" as the keyword and collected 766 articles. Basic information (title, authors, institutions, publisher, publication year, and keywords) about the literature was obtained by web crawlers. Based on the literature, preprocessing of paragraph filtering and text editing was conducted to build the corpus database of the knowledge source. Then, some corpus was annotated manually; based on the annotated corpus, the dataset was generated by data augmentation, and the training set and validation set were divided according to the ratio of 7:3. The dataset information is shown in Table 3.

Table 3. Dataset information.

Data Type	Training Set	Validation Set
Sentences count	4864	2084
Characters count	78,392	30,484

Based on the TensorFlow deep learning framework, the BiLSTM-CRF knowledge extraction model was constructed, and the datasets were input for model training. The parameter settings of the BiLSTM-CRF model are shown in Table 4. The hidden layer dimension of the LSTM network is 100, and the input eigenvector dimension is correspondingly 100. Every 120 sentences are taken as a batch. The global learning rate is set as 0.001, and the Adam optimizer is used. The dropout rate is set to 0.5 to prevent the model from overfitting. Additionally, to avoid the influence of sentence order on model training, sentences in each batch were shuffled before input.

Table 4. Parameter settings of the BiLSTM-CRF model.

Parameters	Value
Hidden layer dimensions	100
Eigenvector dimensions	100
Batch size	120
Learning rate	0.001
Optimizer	Adam
Dropout	0.5

4.2. Accuracy Evaluation

For the accuracy evaluation of the knowledge recognition and extraction, combining the accuracy rate and recall rate, the F score is used to evaluate the effect of the BiLSTM-CRF model.

$$\mathrm{F} = \frac{\left(\beta^2 + 1\right) precision \times recall}{\left(\beta^2 \times precision\right) + recall} \quad (5)$$

In the formula, the precision represents the proportion of correctly recognized labels in the result, which is used to measure the accuracy of the recognition. The recall represents the proportion of correctly recognized labels in the total number of this type of label, which reflects the comprehensiveness of knowledge recognition. β determines the importance of precision and recall. In this study, β was set to 1, which means that the precision and recall were considered to have the same importance. The knowledge recognition result evaluation is shown in Table 5.

Table 5. Accuracy evaluation of knowledge recognition based on BiLSTM-CRF.

Precision	Recall	F
93.64%	84.31%	88.73

The precision shows that the model can obtain accurate recognition results, but the recall rate is relatively lower, which means that the model's knowledge recognition

comprehensiveness is slightly poor. The comprehensive evaluation result of the F score is 88.73, which confirms the feasibility of the BiLSTM-CRF model's knowledge extraction.

Table 6 shows the recognition accuracy evaluation of each type of knowledge entity. Overall, the F scores of all these entities are above 80, and the "AreaV", "Time", and "Val" have the best recognition accuracy, which proves that the BiLSTM-CRF model has a balanced performance in the recognition and extraction of different knowledge entities.

Table 6. Accuracy evaluation of each knowledge entity.

Type	Precision	Recall	F
Area	94.65%	75.27%	83.86
AreaV	97.32%	88.18%	92.53
Data	91.85%	80.91%	86.04
DataV	95.02%	85.47%	89.99
Parm	90.02%	80.89%	85.21
ParmV	92.45%	83.64%	87.82
Time	97.57%	87.65%	92.35
TimeV	93.24%	85.11%	88.99
Val	96.57%	86.25%	91.12
ValV	90.34%	82.36%	86.17

4.3. Results

Through the pretrained BiLSTM-CRF model, the knowledge in the corpus database was recognized and extracted. After relation association and knowledge fusion, the knowledge entities and relations were exported into Neo4j, and the meteorological simulation knowledge graph was constructed. The statistics of the count of knowledge nodes and edges in the knowledge graph are shown in Table 7.

Table 7. Statistics of nodes and relationships in the meteorological simulation knowledge graph.

Name	Type	Count
LiteratureInformation	Node	3134
InputData	Node	35
SimulationScope	Node	291
ParameterScheme	Node	87
SimulationTime	Node	122
ResultValidation	Node	75
Information	Edge	4032
Input	Edge	230
ScopeSetting	Edge	1045
ParameterSetting	Edge	1146
TimeSetting	Edge	579
Validate	Edge	367

Part of the nodes and relationships of the meteorological simulation knowledge graph are shown in Figure 6. It clearly shows the entities, attributes, concepts, and the relationships between them of three articles (the larger green nodes in the center). The LiteratureInformation nodes show the authors, institutions, publisher, publication year, and keywords of each article. The SimulationScope nodes display the study area, resolution, grid count, grid nested scheme, and vertical layering. The InputData nodes show the data used in the simulation experiment, such as National Centers for Environmental Prediction/Final Operational Global (NCEP/FNL) reanalysis, Moderate Resolution Imaging Spectroradiometer (MODIS) 30° lattice data, European Centre for Medium-Range Weather Forecasts Re-Analysis (ERA) interim reanalysis, and terrain height. The ParameterScheme nodes show the parameterization scheme selections of the simulation experiment, including the microphysics scheme, radiation scheme, and cumulus convection scheme. SimulationTime nodes refer to the simulation time settings, such as the simulation start

time, end time, integration time, and integration interval. The ResultValidation nodes show the evaluations of simulation results, including the indicators of mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and correlation coefficient (R). The edges contact nodes show the semantic relationships between these nodes. Using these six types of nodes and relationships can describe the attributes and concepts of the meteorological simulation research clearly and accurately.

Figure 6. Instance of meteorological simulation knowledge graph (part).

As the graph structure can clearly express complex the interconnected relationships between entities, the entities and relations of meteorological simulation knowledge can be retrieved by the Cypher language, which is a query language provided by Neo4j. Users can obtain feasible knowledge related to their research and study efficiently by specific conditional queries. Figure 7 shows some examples for knowledge retrieval.

Figure 7. Knowledge retrieval by Cypher query language, (**a**) input data query; (**b**) study area query; (**c**) parameter setting query; (**d**) JSON file of parameter setting query result.

Figure 7a shows the result of the input data used in the simulation process retrieved by literature name. The Cypher query sentence is as follows:

match(n: title {name: 'WRF land use/cover data optimization and its application in urban heat island simulation(in Chinese)'})<-[r: Input]-(data) return n, data.

The retrieval result shows that the land use/cover data of MODIS2001 and MODIS2012, the initial field and lateral boundary condition of NCEP/FNL reanalysis, and satellite inversion data of GLC2009 and GLC2000 were used in the research of this article to study the impact of land use and land cover on the urban heat island effect.

In Figure 7b, the literature whose study area was Beijing was retrieved by the following Cypher query sentence (to show the results clearly, the number of results is limited to 3).

match(n: title)-[r: ScopeSetting]->(m: studyarea) where m.name = 'Beijing' return n, m limit 3.

Figure 7c shows the retrieval result of parameter selection of the article with the study area restriction. The Cypher query sentence is as follows:

match(n: title)-[r: ScopeSetting]->(m: studyarea{name: 'Beijing'}), (n)-[p: Parameter Setting]->(q) return q, n, m.

The retrieval result can be exported to JSON files, as shown in Figure 7d. JSON is a lightweight data format that is easy for generating and parsing, which can improve the data transmission efficiency. Therefore, it is easy to import the retrieval result into other programs and conduct further research.

In addition to the knowledge retrieval, the association analysis of entities and relationships can be conducted based on the meteorological simulation knowledge graph, which can mine deep information and knowledge. Figure 8 shows the correlation analysis of path discovery based on the meteorological simulation knowledge graph.

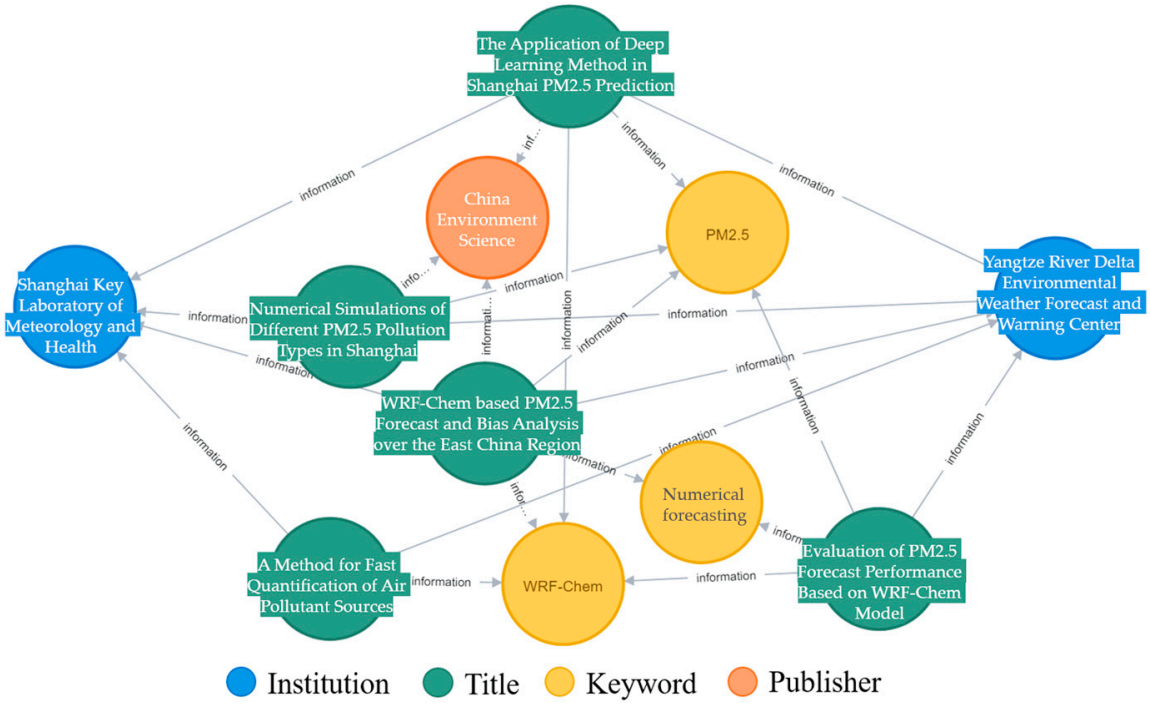

Figure 8. Path discovery between institution entities.

Path discovery refers to querying the shortest or other specific paths of entities based on the relationship network of the knowledge graph. In Figure 8, the path between the Shanghai Key Laboratory of Meteorology and Health and the Yangtze River Delta Environmental Weather Forecast and Warning Center was detected. The detection result shows that the two institutions have published multiple articles collaboratively, and they have conducted in-depth research on the WRF-Chem model, PM2.5, and numerical forecasting. Moreover, it shows the closeness of cooperation between the two institutions. Through the path discovery method of the knowledge graph, the inner relationships and associations between entities can be revealed, which can help users explore the relationships at a deep level and identify new understandings and new knowledge. When the entities and relationships tend to be complex, the superiority of the knowledge graph in data management tends to be more obvious.

To summarize, in this study, 766 articles within the meteorological simulation literature were collected to construct a meteorological simulation knowledge graph; the information on the literature web pages was obtained by web crawler, the knowledge in literature content was extracted by the BiLSTM-CRF model, and the final F score of the model is 88.73. Subsequently, the knowledge graph was constructed by Neo4j; it contains 3744 knowledge entities and 7399 relation edges of meteorological simulation. Efficient retrieval for knowledge and relation, association analysis based on the knowledge graph, and Cypher query language are all demonstrated. The above results proved the feasibility of

the proposed method in meteorological simulation knowledge extraction and knowledge graph construction.

5. Discussion

As knowledge graphs have become a hotspot in knowledge engineering, research on knowledge extraction and knowledge graph construction has been conducted in many fields; however, due to the heterogeneity between different fields, the knowledge extraction and knowledge graph construction methods are not applicable in the meteorological simulation field. There is some research on meteorological knowledge graph construction from structured or semi-structured information on web pages, but the research on knowledge graph constructed from unstructured literature plaintexts is still scarce. In this research, meteorological simulation knowledge in unstructured literature content and semi-structured web pages is extracted, and the meteorological simulation knowledge graph is constructed, which can improve the efficiency of knowledge management, sharing, and reuse for meteorological simulation researchers. However, there are still some deficiencies in the study that need to be improved:

(1). Due to the lack of meteorological simulation corpus, there are still some shortcomings in the dataset constructed by manual annotation and data augmentation, such as corpus inadequacy and annotation error, which may lead to some negative impacts on BiLSTM-CRF model training and knowledge recognition and extraction.
(2). The text features in the literature are not fully utilized; the text features used in this study are the characters and words, but there are other features such as part of speech and word formation that are not used in the text feature processes. In addition, only part of the knowledge was extracted in the literature; the abundant knowledge contained in the whole literature was not extracted sufficiently.
(3). At present, the meteorological simulation knowledge stored in the graph is still insufficient, and the quantity of literature needs to be increased in future work to expand the amount and coverage of knowledge in the meteorological simulation knowledge graph.

6. Conclusions

With the maturity of the meteorological simulation research, the published literature is undergoing a rapid increase, and it becomes more inefficient to obtain specific required knowledge from massive literature. Hence, it is necessary to extract the meteorological simulation knowledge from literature and construct a knowledge graph to transform the unstructured literature content into structured knowledge, and enhance the ability of knowledge management, sharing, and reuse. Knowledge graph research in the meteorological field is still in its infancy, which is due to the lack of training corpus and applicable knowledge extraction methods. Thus, at present, the meteorological simulation knowledge graph is still rare. In this paper, the semi-structured information on the web pages and unstructured knowledge in meteorological simulation literature was extracted by the web crawler and the BiLSTM-CRF model based on natural language processing and deep learning technology. Then, the meteorological simulation knowledge graph was constructed by Neo4j. In addition, efficient knowledge management, knowledge retrieval, and association analysis methods based on the knowledge graph are demonstrated. The proposed methods have universality and are applicable to other fields by reasonable corpus preparation.

The knowledge graph realizes the extraction and structured storage of meteorological simulation knowledge distributed in the literature, and it realizes the efficient management, sharing, and reuse of knowledge. It can compensate for the shortcomings of data-to-knowledge conversion insufficiency and promote the transformation of data services and information services to knowledge services in the meteorological field. Knowledge graph users can conveniently obtain the knowledge related to their research, which can guide their meteorological simulation experiment and assist in complex analysis or decision support. Moreover, users can browse the knowledge of meteorological simulation at the conceptual

level, discover potential connections between entities, strengthen the understanding of complex meteorological processes and simulation experiments, and further enhance the scientificity of meteorological simulation research.

As the research on knowledge graphs is an ongoing hotspot in geographical knowledge engineering, future work could focus on these aspects:

(1). Based on the meteorological simulation knowledge graph, the intelligent Q&A and knowledge recommendation system should be constructed to meet the knowledge needs of multilevel users and enhance the knowledge service capability of the knowledge graph.
(2). Based on the graph mining algorithm, further knowledge reasoning analysis should be conducted to mine more in-depth knowledge from the associated entities and concepts to continuously enrich the knowledge graph, ultimately improving the meteorological simulation knowledge system and promoting the development of meteorological simulation research methods.
(3). Research on knowledge graphs and meteorological simulation models and software should be coupled to develop an automatic meteorological simulation experiment design mechanism, so as to promote the intelligence in meteorological simulation research.

Author Contributions: Z.X. designed and implemented methodology and writing—original draft and editing; and C.Z. proposed the conceptualization and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities, grant number 2652018082.

Acknowledgments: This work was supported by Xinqi Zheng for his suggestions in methodology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bei, N.; Li, X.; Tie, X.; Zhao, L.; Li, G. Impact of Synoptic Patterns and Meteorological Elements on the Wintertime Haze in the Beijing-Tianjin-Hebei Region, China from 2013 to 2017. *Sci. Total Environ.* **2019**, *704*, 135–210. [CrossRef]
2. Alizadeh, R.; Maknoon, R.; Majidpour, M. Clean Development Mechanism, a bridge to mitigate the Greenhouse Gasses: Is it broken in Iran? In Proceedings of the 13th International Conference on Clean Energy, Istanbul, Turkey, 8–12 June 2014.
3. Frederick, A.; David, Y. A Systems Dynamics Approach to Explore Traffic Congestion and Air Pollution Link in the City of Accra, Ghana. *Sustainability* **2010**, *2*, 252–265.
4. Alizadeh, R.; Beiragh, R.G.; Soltanisehat, L.; Soltanzadeh, E. Performance evaluation of complex electricity generation systems: A dynamic network-based data envelopment analysis approach. *Energy Econ.* **2020**, *91*, 104894. [CrossRef]
5. Chen, R.; Zhang, W.; Wang, X. Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere* **2020**, *11*, 676. [CrossRef]
6. Thompson, J.L.; Peterson, T.R. Mediated Modeling: Using Collaborative Processes to Integrate Scientist and Stakeholder Knowledge about Greenhouse Gas Emissions in an Urban Ecosystem. *Soc. Nat. Resour.* **2010**, *23*, 742–757. [CrossRef]
7. Zhang, C.; Chen, M.; Li, R.; Ding, Y.; Lin, H. A virtual geographic environment system for multiscale air quality analysis and decision making: A case study of SO_2 concentration simulation. *Appl. Geogr.* **2015**, *63*, 326–336. [CrossRef]
8. Deng, L.; Liu, Y. Deep learning in knowledge graph. In *Deep Learning in Natural Language Processing*; Springer: Singapore, 2018; pp. 117–145.
9. Zhu, J.; Shi, Q.; Chen, F. Research status and development trends of remote sensing big data. *J. Image Graph.* **2016**, *21*, 1425–1439.
10. Liu, J.; Li, Y.; Duan, H. Knowledge Graph Construction Techniques. *J. Comput. Res. Dev.* **2016**, *53*, 582–600.
11. Xu, Z.; Sheng, Y.; He, L.; Wang, Y.F. Review on Knowledge Graph Techniques. *J. Univ. Electron. Sci. Technol. China* **2016**, *45*, 589–606.
12. Paulheim, H. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semant. Web* **2017**, *8*, 489–508. [CrossRef]
13. Chen, X.; Jia, S.; Xiang, Y. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Appl.* **2019**, *141*, 112948. [CrossRef]
14. Yan, C.; Miao, S.; Liu, Y.; Cui, G. Multiscale modeling of the atmospheric environment over a forest canopy. *Sci. China Earth Sci.* **2020**, *63*, 875–890. [CrossRef]
15. Cheng, C.; Wang, X.; Miao, S.; Wang, Y. Study of Different Methods for Simulating the Impact of Urban Surface on Meteorological Elements in Beijing in Winter. *Plat. Meteorol.* **2014**, *33*, 1045–1056.

16. Li, J.; Zhang, C.; Zheng, X.; Chen, Y. Temporal-Spatial Analysis of the Warming Effect of Different Cultivated Land Urbanization of Metropolitan Area in China. *Sci. Rep.* **2020**, *10*, 2760. [CrossRef]
17. Zhang, C.; Lin, H.; Chen, M.; Yang, L. Scale matching of multiscale digital elevation model (DEM) data and the Weather Research and Forecasting (WRF) model: A case study of meteorological simulation in Hong Kong. *Arab. J. Geosci.* **2014**, *7*, 2215–2223. [CrossRef]
18. Alizadeh, R.; Lund, P.D.; Soltanisehat, L. Outlook on biofuels in future studies: A systematic literature review. *Renew. Sustain. Energy Rev.* **2020**, *134*, 110326. [CrossRef]
19. Yang, J.; Shi, P.; Yang, J.; Gong, D.Y. The impact of the urbanization process on rainfall in Beijing: A case study of 7.21 rainstorm. *Acta Geogr. Sin.* **2020**, *1*, 113–125.
20. Tao, H.; Xing, J.; Zhou, H.; Pleim, J.; Ran, L.; Chang, X.; Wang, S.; Chen, F.; Zheng, H.; Li, J. Impacts of improved modeling resolution on the simulation of meteorology, air quality, and human exposure to PM 2.5, O_3 in Beijing, China. *J. Clean Prod.* **2020**, *243*, 13. [CrossRef]
21. Sahoo, B.; Bhaskaran, P.K. Assessment on historical cyclone tracks in the Bay of Bengal, east coast of India. *Int. J. Climatol.* **2016**, *36*, 95–109. [CrossRef]
22. Fu, L.; Xu, Y.; Xu, Z.H.; Wu, B.; Zhao, D. Tree water-use efficiency and growth dynamics in response to climatic and environmental changes in a temperate forest in Beijing, China. *Environ. Int.* **2019**, *134*, 105209. [CrossRef]
23. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Modell. Softw.* **2019**, *119*, 285–304. [CrossRef]
24. Sowa, J.F. *Principles of Semantic Networks: Exploration in the Representation of Knowledge*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1991; pp. 135–157.
25. Gruber, T.R. A Translation Approach to Portable Ontology Specifications. *Knowl. Acquis.* **1993**, *5*, 199–220. [CrossRef]
26. Yuan, G.; Li, H.; Fan, B. Survey on Development of Knowledge Engineering System. *Comput. Technol. Autom.* **2011**, *30*, 138–143.
27. Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; Zhang, W. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In Proceedings of the The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24 August 2014.
28. Nickel, M.; Murphy, K.; Tresp, V.; Gabrilovich, E. A review of relational machine learning for knowledge graphs. *Proc. IEEE* **2015**, *104*, 11–33. [CrossRef]
29. Bollacker, K.; Cook, R.; Tufts, P. Freebase: A Shared Database of Structured General Human Knowledge. In Proceedings of the AAAI-07, Senior Member Papers Track, Vancouver, BC, Canada, 22–26 July 2007; Volume 7, pp. 1962–1963.
30. Sren, A.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. *Min. Data Financ. Appl.* **2007**, *825*, 722–735.
31. Denny, V.; Markus, K. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* **2014**, *57*, 78–85.
32. Hoffart, J.; Suchanek, F.M.; Berberich, K.; Weikum, G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.* **2013**, *194*, 28–61. [CrossRef]
33. Yan, J.; Lv, T.; Yu, Y. Construction and Recommendation of a Water Affair Knowledge Graph. *Sustainability* **2018**, *10*, 3429. [CrossRef]
34. Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; Sontag, D. Learning a Health Knowledge Graph from Electronic Medical Records. *Sci. Rep.* **2017**, *7*, 1–11. [CrossRef]
35. Ko, H.; Witherell, P.; Lu, Y.; Kim, S.; Rosen, D.W. Machine learning and knowledge graph based design rule construction for additive manufacturing. *Addit. Manuf.* **2020**, *37*, 101620. [CrossRef]
36. Heck, L.; Huang, H. Deep learning of knowledge graph embeddings for semantic parsing of Twitter dialogs. In Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, USA, 3–5 December 2014; pp. 597–601.
37. Gong, J.; Gen, J.; Wu, H. Geospatial Knowledge Service: A Review. *Geom. Inf. Sci. Wuhan Univ.* **2014**, *39*, 883–890.
38. Lu, F.; Yu, L.; Qiu, P. On Geographic Knowledge Graph. *J. Geo Inf. Sci.* **2017**, *19*, 723–734.
39. Xu, F.; Li, H.; Li, X. Named entity recognition in the domain of geographical subject. In Proceedings of the 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, China, 29–31 July 2017.
40. Guo, C.; Xu, T.; Liu, L. Construction of Knowledge Graph. Based on Geographic Ontology. *Conf. Ser. Earth Environ. Sci.* **2019**, *252*, 052161. [CrossRef]
41. Wang, J.; Lu, F.; Wu, S.; Yu, L. Constructing the Corpus of Geographical Entity Relations Based on Automatic Annotation. *J. Geo-Inf. Sci.* **2018**, *20*, 871–879.
42. Zhang, Y.; Zhu, J.; Zhu, Q.; Xie, Y.; Li, W.; Fu, L.; Zhang, J.; Tan, J. The construction of personalized virtual landslide disaster environments based on knowledge graphs and deep neural networks. *Int. J. Digit. Earth* **2020**, *13*, 1637–1655. [CrossRef]
43. Shi, K.; Gong, C.; Lu, H.; Zhu, Y.; Niu, Z. Wide-grained capsule network with sentence-level feature to detect meteorological event in social network. *Futur. Gener. Comp. Syst.* **2020**, *102*, 323–332. [CrossRef]
44. Zhang, C.; Chen, M.; Li, R.; Fang, C.; Lin, H. What's going on about geo-process modeling in virtual geographic environments (VGEs). *Ecol. Model.* **2016**, *319*, 147–154. [CrossRef]
45. Xu, B.; Lin, H.; Chiu, L.; Hu, Y.; Zhu, J.; Hu, M.; Cui, W. Collaborative virtual geographic environments: A case study of air pollution simulation. *Inf. Sci.* **2011**, *181*, 2231–2246. [CrossRef]

46. Lin, H.; Chen, M.; Lu, G.; Zhu, Q.; Gong, J.; You, X.; Wen, Y.; Xu, B.; Hu, M. Virtual Geographic Environments (VGEs): A New Generation of Geographic Analysis Tool. *Earth Sci. Rev.* **2013**, *126*, 74–84. [CrossRef]
47. Chen, J.; Liu, W.; Wu, H.; Li, Z.; Zhang, L. Basic Issues and Research Agenda of Geospatial Knowledge Service. *Geom. Inf. Sci. Wuhan Univ.* **2019**, *44*, 38–47.
48. Rong, X. word2vec Parameter Learning Explained. *arXiv* **2014**, arXiv:1411.2738.
49. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics* **2017**, *34*, 1381–1388. [CrossRef]
50. Sak, H.; Senior, A.; Beaufays, F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv* **2014**, arXiv:1402.1128.
51. Xie, T.; Yang, J.; Liu, H. Chinese Entity Recognition Based on BERT-BiLSTM-CRF Model. *Comput. Syst. Appl.* **2020**, *29*, 48–55.
52. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991.
53. Chen, T.Y. Research on Optimization of Linear Conditional Random Field Training Algorithm. Master's Thesis, Fudan University, Shanghai China, 29 April 2010.
54. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv* **2019**, arXiv:1901.11196.
55. Cui, B.; Gao, J.; Tong, Y.; Xu, J.; Zhang, D.; Zou, L. Progress and Trend in Novel Data Management System. *J. Softw.* **2019**, *30*, 164–193.
56. Mario, M.; Fabio, M.; Mirko, C.; Vincenzo, M.; Antonio, P. GraphDBLP: A System for Analyzing Networks of Computer Scientists Through Graph Databases. *Multimed. Tools Appl.* **2018**, *77*, 18657–18688.
57. Zhang, L.; Xiong, S. Design and Implementation of Social Network Platform Based on Neo4j. *Inf. Res.* **2018**, *250*, 81–86.
58. Holzschuher, F.; Peinl, R. Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4j. In Proceedings of the ACM International Conference Proceeding Series, Genua, Italy, 18 March 2013. [CrossRef]